PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-161340

(43)Date of publication of application : 21.06.1996

(51)Int.Cl. G06F 17/28
G06F 17/22
G06F 17/27

(21)Application number : 06-307223

(22)Date of filing : 12.12.1994

(71)Applicant : RICOH CO LTD

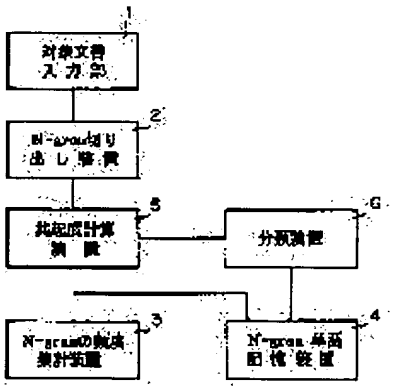
(72)Inventor : KATOOKA TAKASHI

(54) AUTOMATIC COMPOUND WORD EXTRACTION DEVICE

(57)Abstract:

PURPOSE: To efficiently and automatically collect a compound word whose degree of cooccurrence is large with respect to an idiom and a compound word by the combination of words whose speciality is not high.

CONSTITUTION: An N-gram segment device 2 segments N-gram of a word from an objective document which is read from an objective document input part 1. A frequency adding-up device 3 adds up the appearing frequency of the compound word of segmented N-gram and a word storage device 4 stores the N- gram compound word and appearing frequency that the frequency adding-up device 3 adds up. A cooccurrence degree calculation device 5 calculates the cooccurrence degree of N-gram by using appearing frequency in the objective document of the respective words constituting N-gram and the appearing frequency of N-gram itself. A classification device 6 rearranges information in the word storage device 4 by the value of the cooccurrence degree calculated by the cooccurrence degree calculation device 5.



LEGAL STATUS

[Date of request for examination] 16.11.2000

[Date of sending the examiner's decision of rejection] 20.08.2002

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－161340

(43)公開日　平成8年(1996)6月21日

(51)Int.Cl.⁶	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ 17/28					
17/22					
17/27					
		8420－5Ｌ	Ｇ０６Ｆ 15/38	Ｃ	
		9288－5Ｌ	15/20	５１４ Ｕ	

審査請求　未請求　請求項の数2　ＯＬ　（全 7 頁）　最終頁に続く

(21)出願番号　特願平6－307223

(22)出願日　平成6年(1994)12月12日

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号

(72)発明者　加登岡　隆
東京都大田区中馬込１丁目３番６号　株式会社リコー内

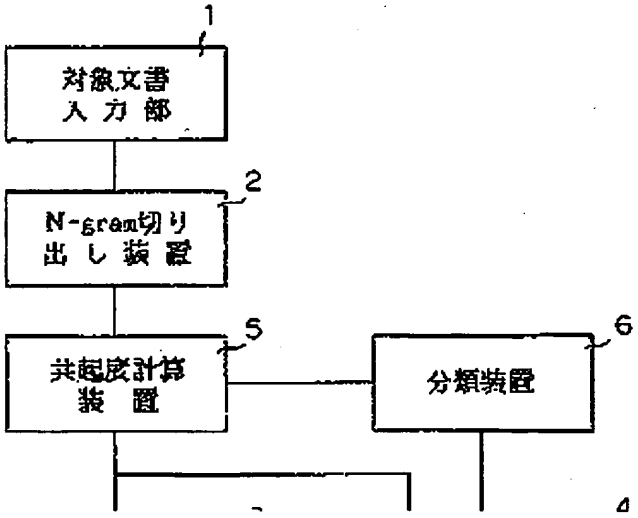
(74)代理人　弁理士　高野　明近

(54)【発明の名称】　連語自動抽出装置

(57)【要約】

【目的】　専門性の高い語でない語の組合せによる熟語や連語に対しても、効率良く、共起の度合いが強い連語を自動的に収集する。

【構成】　対象文書入力部１から読み込まれた対象文書から単語のN-gramをN-gram切り出し装置２により切り出す。切り出されたN-gramの連接語の出現頻度を頻度集計装置３により集計し、N-gram連接語と、前記頻度集計装置３により集計された出現頻度を単語記憶装置４により記憶する。前記N-gramを構成する各単語の対象文書中の出現頻度と、N-gram自体の出現頻度とを用いてN-gramの共起度を共起度計算装置５により計算する。分類装置６は前記共起度計算装置５で計算された共



1

【特許請求の範囲】

【請求項１】　対象文書を読み込む対象文書入力部と、該対象文書入力部に読み込まれた対象文書から単語のN-gram（N＝１，２，３，…，Nmax）を切り出す切り出し装置と、該切り出し装置により切り出されたN-gramの連接語の出現頻度を集計する頻度集計装置と、N-gram連接語と該連接語の出現頻度を記憶する単語記憶装置と、前記N-gramを構成する各単語の対象文書中の出現頻度とN-gram自体の出現頻度とを用いてN-gramの共起度を計算する共起度計算装置と、該共起度計算装置により計算された共起度の値で前記単語記憶装置内の情報を並べ変える分類装置とを有することを特徴とする連語自動抽出装置。

【請求項２】　用語として抽出すべき条件に合わない構成を記憶する条件設定記憶装置と、前記単語記憶装置中に記憶した構成に合うN-gramを排除するためのパターンマッチ装置とを用いてさらに精度良く連語を抽出することを特徴とする請求項１記載の連語自動抽出装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、言語処理装置における連語自動抽出装置に関し、より詳細には、対象文書から連語を効率よく自動収集するための連語自動抽出装置に関する。例えば、機械翻訳やワードプロセッサなどの用語辞書作成装置に適用されるものである。

【０００２】

【従来の技術】従来の言語処理装置について記載した公知文献としては、例えば、特開平６－１９９６８号公報がある。この公報のものは、膨大な単語のなかから専門用語を容易に抽出できるようにし、専門用語辞書の構築を短時間で容易に行うことができるようにするために、単語分割装置で入力文が単語に区切られて品詞情報が付与される等の正規化が行われ、単語分割装置で正規化された入力データは専門用語判定装置に出力され、この専門用語判定装置で前記各辞書を参照しながら各単語の評価が行われ、この評価に従って専門用語の候補が抽出されるものである。しかし、専門用語判定は、構成語数、構成語の使用頻度、分野別用語辞書、字種（カタカナ語）を考慮して行い、分野別用語辞書を必要としている。また、前記公報のものは、判定対象の専門用語候補選定についての記述がない。

【０００３】

【発明が解決しようとする課題】例えば、機械翻訳シス

2

おり、装置が重くなるという欠点があった。

【０００４】本発明は、このような実情に鑑みてなされたもので、専門性の高い語でない語の組合せによる熟語や連語に対しても、複数語からなるエントリを共起しているのか、あるいは偶然による連接かを見極め、効率良く、共起の度合いが強い連語を自動的に収集するようにした連語自動抽出装置を提供することを目的としている。

【０００５】

【課題を解決するための手段】本発明は、上記課題を解決するために、（１）対象文書を読み込む対象文書入力部と、該対象文書入力部に読み込まれた対象文書から単語のN-gram（N＝１，２，３，…，Nmax）を切り出す切り出し装置と、該切り出し装置により切り出されたN-gramの連接語の出現頻度を集計する頻度集計装置と、N-gram連接語と該連接語の出現頻度を記憶する単語記憶装置と、前記N-gramを構成する各単語の対象文書中の出現頻度（N＝１の場合）とN-gram自体の出現頻度とを用いてN-gramの共起度を計算する共起度計算装置と、該共起度計算装置により計算された共起度の値で前記単語記憶装置内の情報を並べ変える分類装置とを有すること、更には、（２）用語として抽出すべき条件に合わない構成を記憶する条件設定記憶装置と、前記単語記憶装置中に記憶した構成に合うN-gramを排除するためのパターンマッチ装置とを用いてさらに精度良く連語を抽出することを特徴としたものである。

【０００６】

【作用】前記構成を有する本発明の連語自動抽出装置は、（１）対象文書を入力し、該対象文書から単語のN-gram（N＝１，２，３，…，Nmax）を切り出し、切り出されたN-gramの連接語の出現頻度を集計し、N-gram連接語とその出現頻度を記憶し、前記N-gramを構成する各単語の対象文書中の出現頻度（N＝１の場合）とN-gram自体の出現頻度とを用いてN-gramの共起度を計算し、共起度の値で記憶装置内の情報を並べ変えることにより、連語の構成語の共起の強さをその構成語の出現頻度と連語の出現頻度から求めることができるので、辞書などを使用しないで、簡単な装置で、入力文において、強い共起をもって連語として出現する語を効率良く自動的に迅速に抽出することができる。（２）用語として抽出すべき条件に合わない構成を記憶し、記憶した構成に合うN-gramを排除するためのパターンマッチ装置を用いてさらに精度良く連語を抽出するので、前記（１）で

する。まず、対象文書からN-gramの連接語を抽出する（N＝1，2，3，…，Nmax)。対象文書が英語の場合であれば、言語の形態的特徴からスペース文字等を参考にして一語一語を分割する。N＝3の3-gramであれば、最大3単語連接する用語を切り出す。抽出したN連接語は、出現頻度を集計する装置により、その出現頻度をカウント集計する。また、各一単語ごとの出現頻度をカウントして集計する。この結果は、N-gram単語記憶装置に記憶される。

【0008】入力文書に対する出現頻度の集計が終わる＊1

＊と、N-gram単語記憶装置内のN-gramの連接語に対して共起度の計算を、例えば、以下の式に従って計算する。N連接の語が、つまり連語の構成語がそれぞれw1，w2，w3，…，wNの時、それぞれの出現頻度がH(w1)，H(w2)，…，H(w3)で、N連接語自体の出現頻度がH(w1，w2，w3，…，wN)と表す。また、対象入力文書の総語数をAとする。

【0009】

【数1】

$$\frac{P(w1, w2, w3, \ldots, wN)}{P(w1)\times P(w2)\times P(w3)\times \ldots \times P(wN)} \quad \cdots (1)$$

ここで、

$$P(w1)=\frac{H(w1)}{A}$$

$$P(w2)=\frac{H(w2)}{A}$$

$$P(w3)=\frac{H(w3)}{A}$$

$$\cdots\cdots$$

$$P(wN)=\frac{H(wN)}{A}$$

$$P(w1, w2, w3, \ldots, wN) = \frac{H(w1,w2,w3,\ldots,wN)}{A-(N-1)} \quad \cdots (2)$$

である。

Aの値がNに比べて充分大きいとき、式（2）は $\frac{H(w1,w2,w3,\ldots,wN)}{A}$ と近似できる。

【0010】（1）式の分母は、連語を構成する各単語の出現確率から各語が偶然に連接する確率を表す。（1）式の分子は、実際に各語が連接して出現する確率である。したがって、（1）式はある連語が実際に連接する確率と偶然に連接する確率との比となる。（1）式の値が高いほど、そのN-gramの連接語は、共起して出現する度合いが高いといえる。逆に低い場合は、共起するよりも偶然に連接したものである可能性が高い。

【0011】図1は、本発明による連語自動抽出装置の一実施例（実施例1）を説明するための構成図で、図中、1は対象文書入力部、2はN-gram切り出し装置、3はN-gramの頻度集計装置、4はN-gram単語記憶装置

頻度集計装置3により集計し、N-gram連接語と、前記頻度集計装置3により集計された出現頻度を単語記憶装置4により記憶する。

【0013】前記N-gramを構成する各単語の対象文書中の出現頻度（N＝1の場合）と、N-gram自体の出現頻度とを用いてN-gramの共起度を共起度計算装置5により計算する。分類装置6は前記共起度計算装置5で計算された共起度の値で単語記憶装置4内の情報を並べ変える。このようにして、専門性の高い語でない語の組合せによる熟語や連語に対しても、複数語からなるエントリを共起しているか、あるいは偶然による連接かを見きわめ、効率の良い、共起の度合いが強い連語を自動的に

部１より原文の先頭からｉ単語目からｊ単語を入力し、変数wordsに格納する（Ｓ３）。次に、（ｉ＋ｊ－１）番目の単語が存在するかどうかを判断し（Ｓ４）．単語が存在していれば、次に．出現頻度記憶装置４にwordsの中の単語列が既に存在するかどうかを判断する（Ｓ５）。存在しなければ、wordsの中味を出現頻度記憶装置４に出現回数１として記憶し（Ｓ６）、変数ｊを１だけカウントアップし（Ｓ７）、前記ステップＳ２へ戻る。

【００１５】前記ステップＳ５において、出現頻度記憶装置４にwordsの中の単語列が既に存在していれば、出現頻度記憶装置４に記憶されているwordsの中味の出現回数を１だけカウントアップし（Ｓ８）、前記ステップＳ７へ行く。前記ステップＳ２において、変数ｊの値が最大追語連接数Ｎの値を越えたら、ｊに１をセットし、ｉを１カウントアップし（Ｓ９）、前記ステップＳ３へ＊＊行く。前記ステップＳ４において、（ｉ＋ｊ－１）番目の単語が存在しなければ．次に、ｊがｉと等しいかどうかを判断し（Ｓ１０）、等しくなければ、前記ステップＳ９へ行き、等しければ．対象文書の総語数を記憶する変数Ａに，－１をセットする（Ｓ１１）。

【００１６】次に、出現頻度記憶装置４に記憶された登大Ｎ個の連語の共起度を計算する。結果を出現頻度記憶装置４に記憶し（Ｓ１２）．分類装置６により出現頻度記憶装置４に記憶された情報を共起度の高い順に並び変える（Ｓ１３）。

【００１７】以下、本発明の実施例１について．例文に基づき具体的に説明する。対象文書入力部から原文を入力する。

【００１８】

【表１】

The orchestra gave him superb support. ←―― 入力文

（語番号1　2　3　4　5　6）

【００１９】最大３gramの追語を自動抽出する実施例について説明する。まず、１変数ｉとｊに初期値１をセットする（Ｓ１）。対象入力部１から対象文書を読み込み．変数ｊが最大連語数３を越えていないので（Ｓ２）．先頭（ｉ＝１）から１単語（ｊ＝１）を得る（Ｓ３）。つまり．"The"が得られる。ｉ＋ｊ－１＝１番目の単語は存在するので（Ｓ４）、この単語がＮ-gram単語記憶装置４に記憶されているかどうか調べる（Ｓ５）。まだ記憶されていないので、新規にＮ-gram単語記憶装置４に語"The"をその出現回数１として記憶する（Ｓ６）。既に記憶されてる場合は出現回数を１だけカウントアップする（Ｓ８）。ｊを１だけカウントアップし（Ｓ７）．次に先頭（ｉ＝１）から２単語（ｊ＋１）を得る。

【００２０】つまり、"The orchestra"を得る。ｉ＋ｊ－１＝２番目の単語は存在するので（Ｓ４）．この単語がＮ-gram単語記憶装置４に記憶されているかどうか調べる。まだ記憶されていないので．新規に記憶し出現回数を１とする。同様にｊ＝３のときは、"The orchestra gave"を得る。ｊを１つカウントアップすると（Ｓ７）．ｊ＝４となる。ｊが最大連語追接数３を越えるので．ｊに１をセットしｉを１カウントアップして２とする（Ｓ９）。

【００２１】次に、第２番目（ｉ＝２）の単語から１単語（ｉ＝１）を得る。"orchestra"が得られる。ｉ＝

されている。この後、式（１）に従って出現頻度から各連語の共起度を計算し、結果を単語記憶装置４に記憶する（Ｓ１２）。さらに、共起度の値の高い順に分類（ソート）し直す。

【００２２】例えば、対象入力文を１００万語用意し、最大３の連接語を抽出する場合、共起度計算後のＮ-gram単語記憶装置４は．図５のようになる。Ｎ＝１の場合については共起度は求めていない。共起度の値の大きさでソートし、例えば、上位５０％を候補とするなどのしきい値を決めたり、連語の数が同じものごとに共起度の値でソート（ソート方法としては、クイックソート、マージソート、単純ソート法などがある）し、それぞれの長さの連語ごとに抽出する追語の共起度の値をあるしきい値以上と設定するというやり方で共起度の強い追語を自動的に抽出することができる。

【００２３】図６は、本発明による追語自動抽出装置の他の実施例（実施例２）を説明するための構成図で、図中．７はパターンマッチ装置、８は条件設定記憶装置で．その他、図１と同じ作用をする部分は同一の符号を付してある。

【００２４】対象文書を対象文書入力部１から読み込み．該対象文書入力部１から読み込まれた対象文書から単語のＮ-gram（Ｎ＝１，２，３，…，Ｎmax）をＮ-gram切り出し装置２により切り出す。該Ｎ-gram切り出し装置２により切り出されたＮ-gramの連接語の出現頻度を

算された共起度の値で単語記憶装置４内の情報を並べ変える。このようにして、専門性の高い語でない語の組合せによる熟語や連語に対しても、複数語からなるエントリを共起しているか、あるいは偶然による連接かを見きわめ、効率の良い、共起の度合いが強い連語を自動的に収集することができる。

【００２６】条件設定記憶装置８は、用語として抽出すべき条件に合わない構成を記憶し、パターンマッチ装置７は、前記条件設定記憶装置８中に記憶した構成に合うN-gramを排除するためのもので、これらを用いることにより、さらに程度良く連語を抽出することができる。すなわち、実施例２においては、用語として抽出条件に合わない構成とは、連語となりにくいパターンをさしている。例えば、冠詞とある１単語の２連接語の場合、あるいは代名詞（英語では、his,my,your,their,them,himなど）と別の語の２連接などがある。

【００２７】実施例２の例を以下に示す。表２の例は、連接数を２と限定し、その場合の２連接語の第一語が装置内に記憶された"the"，"a"，"an"，"his"，…などのいずれかであるとき、それを抽出用語の対象から排除する場合である。パターンマッチはパターンマッチ装置７により文字列の比較により行われる。

【００２８】

【表２】

連接数	先頭語
2	the
2	a
2	an
2	his
2	her
2	their
2	my
2	your

【００２９】また、表３の例では、連接数を２と限定し、その場合の２連接語の第一語が装置内に記憶された"the"，"a"，"an"のいずれかであり、かつ第２の単語が"in"，"of"，"with"，…"from"，"to"のいずれかであるその連語を抽出用語の対象から排除する場合である。パターンマッチはパターンマッチ装置７により文字列の比較により行われる。

【００３０】

【表３】

連接数	先頭語	第２単語
2	the	in
	a	of
	an	with
		for
		on
		from
		to

【００３１】

【発明の効果】以上の説明から明らかなように、本発明によると、以下のような効果がある。

（１）請求項１に対応する効果：連語の構成語の共起の強さをその構成語の出現頻度と連語の出現頻度から求めることができるので、辞書などを使用しないで、簡単な装置で、入力文において、強い共起をもって連語として出現する語を効率良く自動的に迅速に抽出することができる。

（２）請求項２に対応する効果：前記（１）で抽出した語の中で用語として抽出するには不適切であると思われるN-gramパターンを予め条件設定記憶装置に記憶しておくので、用語候補からこのパターンにマッチするものを排除することができ、精度良く用語を抽出することができる。

【図面の簡単な説明】

【図１】　本発明による連語自動抽出装置の一実施例を説明するための構成図である。

【図２】　本発明による連語自動抽出装置の動作を説明するためのフローチャート（その１）である。

【図３】　本発明による連語自動抽出装置の動作を説明するためのフローチャート（その２）である。

【図４】　本発明におけるN-gram単語記憶装置の記憶例を示す図である。

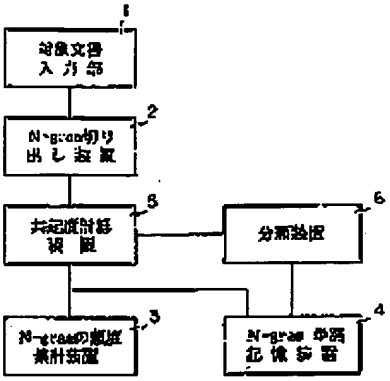
【図５】　本発明におけるN-gram単語記憶装置の他の記憶例を示す図である。

【図６】　本発明による連語自動抽出装置の他の実施例を説明するための構成図である。

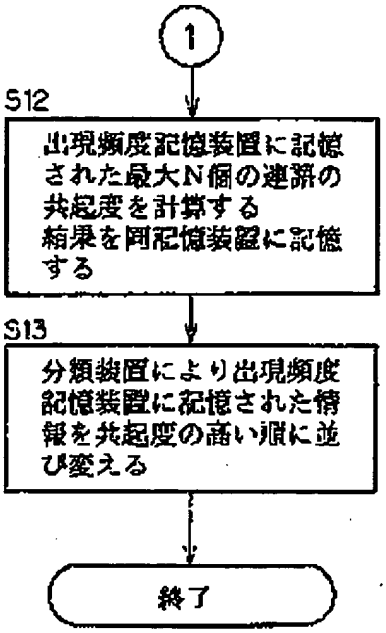
【符号の説明】

１…１は対象文書入力部、２…N-gram切り出し装置、３…N-gramの頻度集計装置、４…N-gram単語記憶装置、５…共起度計算装置、６…分類装置、７…パターンマッチ装置、８…条件設定記憶装置。

【図1】



【図3】



【図4】

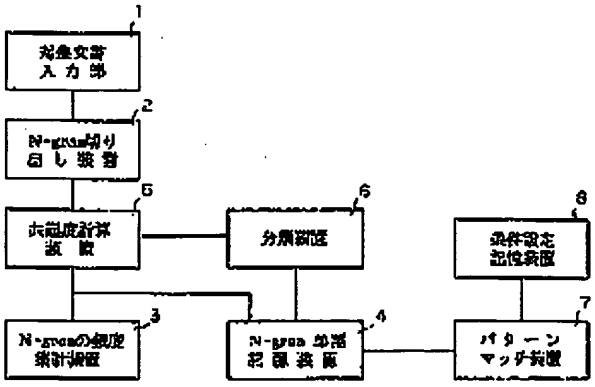
連鎖	出現頻度	共起度
The	1	–
The orchestra	1	–
The orchestra gave	1	–
orchestra	1	–
orchestra gave	1	–
orchestra gave him	1	–
gave	1	–
gave him	1	–
gave him superb	1	–
him	1	–
him superb	1	–
him superb support	1	–
superb	1	–
superb support	1	–
support	1	–

N-gram単語記憶装置の記憶例

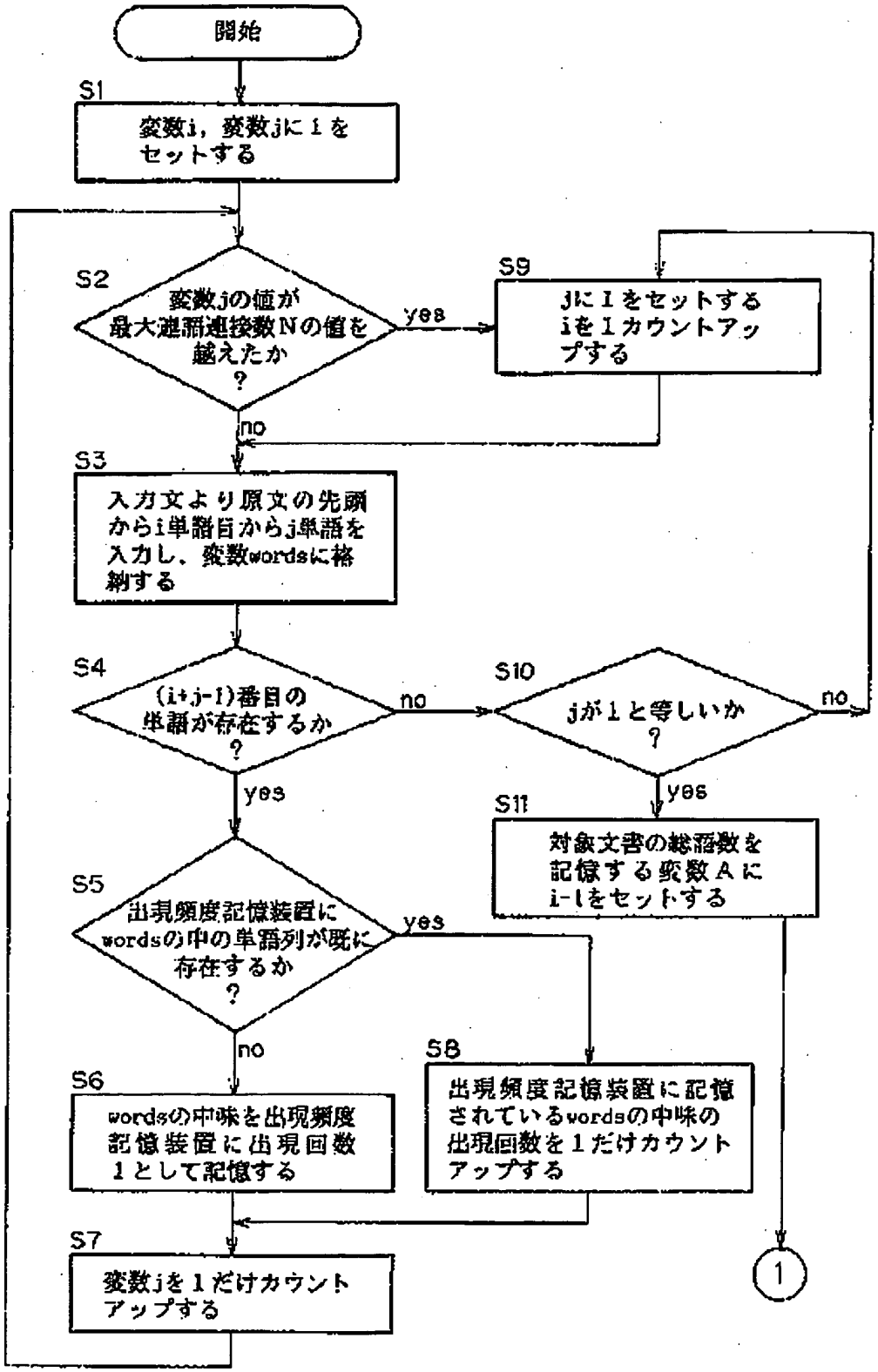
【図5】

N連鎖	出現頻度	共起度
securities brings fannie	3	396920
securities	1654	
brings	64	
fannie	119	
the khmer rouge	12	306784
the	116415	
khmer	14	
rouge	24	
calls seeking comment	10	138985
calls	335	
seeking	342	
comment	628	
premium over yesterday's	6	57825
premium	165	
over	2437	
yesterday's	220	
public employees retirement	6	47793
public	1052	
employees	586	
retirement	193	
securities brings	8	758
brings fannie	8	1050
the khmer	12	7
khmer rouge	14	41657
calls seeking	11	96
seeking comment	11	31
premium over	28	70
over yesterday's	8	8
public employees	10	13
employees retirement	8	69

【図6】



【図2】


開始
S1
変数i，変数jに1をセットする
S2
変数jの値が最大連続連接数Nの値を越えたか？
yes
S9
jに1をセットするiを1カウントアップする
no
S3
入力文より原文の先頭からi単語目からj単語を入力し、変数wordsに格納する
S4
(i+j-1)番目の単語が存在するか？
no
S10
jが1と等しいか？
no
yes
yes
S11
対象文書の総語数を記憶する変数Aにi-1をセットする
S5
出現頻度記憶装置にwordsの中の単語列が既に存在するか？
yes
no
S6
wordsの中味を出現頻度記憶装置に出現回数1として記憶する
S8
出現頻度記憶装置に記憶されているwordsの中味の出現回数を1だけカウントアップする
S7
変数jを1だけカウントアップする
1

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] The object document input section which reads an object document, and the logging equipment which starts N-gram (N= 1, 2 and 3, --, Nmax) of a word from the object document read into this object document input section, The frequency total equipment which totals the frequency of occurrence of the connection word of N-gram started by this logging equipment, The word storage which memorizes the frequency of occurrence of a N-gram connection word and this connection word, Whenever [coincidence / which calculates whenever / coincidence / of N-gram / using the frequency of occurrence in the object document of each word which constitutes said N-gram, and the frequency of occurrence of N-gram itself] Count equipment, Copula automatic extracting equipment characterized by having classification equipment into which the information in said word storage is put in order and changed with the value of whenever [coincidence / which was calculated by count equipment whenever / this coincidence].
[Claim 2] Copula automatic extracting equipment according to claim 1 characterized by extracting a copula with a still more sufficient precision using the pattern match equipment for eliminating N-gram suitable for the configuration memorized in the conditioning storage which memorizes the configuration which does not suit the conditions which should be extracted as vocabulary, and said word storage.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Industrial Application] This invention relates to the copula automatic extracting equipment for carrying out automatic collection of the copula efficiently from an object document at a detail more about the copula automatic extracting equipment in language-processing equipment. For example, it is applied to vocabulary dictionary listing devices, such as machine translation and a word processor.
[0002]
[Description of the Prior Art] As well-known reference which indicated conventional language-processing equipment, there is JP,6-19968,A, for example. In order for the thing of this official report to enable it to extract a technical term easily out of a huge word and to enable it to build a technical-term dictionary easily for a short time An input statement is divided into a word with word division equipment, and normalization of part-of-speech information being given is performed. The input data which it normalized with word division equipment is outputted to technical-term judging equipment, while this technical-term judging equipment refers said each dictionary, evaluation of each word is performed, and the candidate of a technical term is extracted according to this evaluation. However, a technical-term judging is performed in consideration of the number of configuration words, the operating frequency of a configuration word, the vocabulary dictionary classified by field, and a type of letters (katakana word), and the vocabulary dictionary classified by field is needed. Moreover, the thing of said official report does not have the description about the technical-term candidate selection for a judgment.
[0003]
[Problem(s) to be Solved by the Invention] For example, in a machine translation system, in translating the text of a certain specific field, it is, or it registers the vocabulary of the field into the dictionary in advance how much, and the engine performance of a translation acts greatly. However, by the conventional technical-term detection approach of having used unknown word retrieval, there was a fault that the technical term which consists of two or more words could not be efficiently extracted like an idiom or a copula. Moreover, the thing of said official report was collecting vocabulary using the vocabulary dictionary classified by field, and there was a fault that equipment became heavy.
[0004] This invention discerns coinciding the entry which was not made in view of such the actual condition, and consists of two or more words also to the idiom and copula by combination of the word which is not a high word of an expert, or the connection depended by chance, and it is efficient and it aims at offering the copula automatic extracting equipment with which the degree of coincidence collected strong copulas automatically.
[0005]
[Means for Solving the Problem] The object document input section which reads the document for (1) in order that this invention may solve the above-mentioned technical problem, The logging equipment which starts N-gram (N= 1, 2 and 3, --, Nmax) of a word from the object document read into this object document input section, The frequency total equipment which totals the frequency of occurrence of the connection word of N-gram started by this logging equipment, The word storage which memorizes the frequency of occurrence of a N-gram connection word and this connection word, Whenever [coincidence / which calculates whenever / coincidence / of N-gram / using the frequency of occurrence in the object document of each word which constitutes said N-gram (in the case of N= 1), and the frequency of occurrence of N-gram itself] Count equipment, having classification equipment into which the information in said word storage is put in order and changed with the value of whenever [coincidence / which was calculated by count equipment whenever / this coincidence] -- further (2) It is characterized by extracting a copula with a still more sufficient precision using the pattern match equipment for eliminating N-gram suitable for the configuration memorized in the conditioning storage which memorizes the configuration which does not suit the

conditions which should be extracted as vocabulary, and said word storage.

[0006]

[Function] The copula automatic extracting equipment of this invention which has said configuration (1) Input an object document and N-gram (N= 1, 2 and 3, --, Nmax) of a word is started from this object document. Total the frequency of occurrence of the connection word of started N-gram, and a N-gram connection word and its frequency of occurrence are memorized. By calculating whenever [coincidence / of N-gram] using the frequency of occurrence in the object document of each word which constitutes said N-gram (in the case of N= 1), and the frequency of occurrence of N-gram itself, and putting in order and changing the information in storage with the value of whenever [coincidence] The word which appears as a copula with strong coincidence in an input statement with easy equipment can be quickly extracted efficiently automatically without using a dictionary etc., since it can ask for the strength [the configuration word of a copula] of coincidence from the frequency of occurrence of the configuration word, and the frequency of occurrence of a copula. (2) Memorize the configuration which does not suit the conditions which should be extracted as vocabulary, and since a copula is extracted with a still more sufficient precision using the pattern match equipment for eliminating N-gram suitable for the configuration which memorized By memorizing beforehand the N-gram pattern considered to be unsuitable for extracting as vocabulary in the word extracted above (1) to the conditioning store, what matches this pattern from a vocabulary candidate can be eliminated, and the vocabulary can be extracted with a sufficient precision.

[0007]

[Example] An example is explained below with reference to a drawing. First, the connection word of N-gram is extracted from an object document (N= 1, 2 and 3, --, Nmax). If it is the case where an object document is English, it will refer to a space character etc. from the gestalt-description of language, and every word will be divided. If it is 3-gram of N= 3, the vocabulary which carries out a maximum of 3 word connection will be started. Extracted N connection word carries out the count total of the frequency of occurrence with the equipment which totals the frequency of occurrence. Moreover, the frequency of occurrence for every word is counted, and it totals. This result is memorized by N-gram word storage.

[0008] After the total of the frequency of occurrence to an input-statement document finishes, count of whenever [coincidence] is calculated according to the following formulas to the connection word of N-gram in N-gram word storage. When the words of N connection, i.e., the configuration word of a copula, are w1, w2, w3, --, wN, respectively, the frequency of occurrence of the N connection word itself expresses [each frequency of occurrence] with H (w1), H (w2), --, H (w3) H (w1, w2, w3, --, wN). Moreover, the total number of words of an object input-statement document is set to A.

[0009]

[Equation 1]

$$\frac{P(w1, w2, w3, ..., wN)}{P(w1) \times P(w2) \times P(w3) \times ... \times P(wN)} \quad \cdots (1)$$

ここで、

$$P(w1) = \frac{H(w1)}{A}$$

$$P(w2) = \frac{H(w2)}{A}$$

$$P(w3) = \frac{H(w3)}{A}$$

. . . . . .

$$P(wN) = \frac{H(wN)}{A}$$

$$P(w1, w2, w3, ..., wN) = \frac{H(w1,w2,w3,...wN)}{A-(N-1)} \quad \cdots (2)$$

である。

Aの値がNに比べて充分大きいとき、式（2）は $\frac{H(w1,w2,w3,...wN)}{A}$

と近似できる。

[0010] (1) The denominator of a formula expresses the probability which each word connects by chance from the appearance probability of each word which constitutes a copula. (1) The molecule of a formula is the probability for each word to be connected and to actually appear. Therefore, (1) type serves as a ratio of the probability which a certain copula actually connects, and the probability connected [chance]. (1) It can be said that the connection word of the N-gram has the high degree which coincides and appears, so that the value of a formula is high. conversely, when low, it coincides -- as -- ** -- possibility of connecting by chance is high.
[0011] Drawing 1 is a block diagram for explaining one example (example 1) of the copula automatic extracting equipment by this invention, and, for N-gram logging equipment and 3, as for N-gram word storage (frequency-of-occurrence storage) and 5, the frequency total equipment of N-gram and 4 are [one / the object document input section and 2 / count equipment and 6] classification equipment whenever [coincidence] among drawing.
[0012] N-gram (N= 1, 2 and 3, --, Nmax) of a word is started with N-gram logging equipment 2 from the object document which read the object document from the object document input section 1, and was read from this object document input section 1. The frequency of occurrence of the connection word of N-gram started by this N-gram logging equipment 2 is totaled with frequency total equipment 3, and the frequency of occurrence totaled by said frequency total equipment 3 is remembered to be a N-gram connection word with the word storage 4.
[0013] Whenever [coincidence / of N-gram] is calculated with count equipment 5 whenever [coincidence] using the frequency of occurrence in the object document of each word which constitutes said N-gram (in the case of N= 1), and the frequency of occurrence of N-gram itself. Classification equipment 6 puts in order and changes the information in the word storage 4 with the value of whenever [coincidence / which was calculated with count equipment 5 whenever / said coincidence]. Thus, it can discern whether they are whether the entry which consists of two or more words is coincided also to the idiom and copula by combination of the word which is not a high word of an expert, and the connection depended by chance, and the efficient degree of coincidence can collect strong copulas automatically.
[0014] Drawing 2 and drawing 3 are the flow charts for explaining actuation of the copula automatic extracting equipment by this invention. Hereafter, according to each step (S), it explains in order. First, Variable i and Variable j are set to 1 (S1), and it judges whether the value whose value of Variable j is the number N of the maximum copula connection was exceeded (S2). If it is not over the value of N next, from the object document input section 1, j word is inputted from i word eye from the head of the text, and it stores in Variable words (S3). Next, if it judges whether the word of eye watch (i+j -1) exists and (S4) and a word exist next, it will judge whether the word train in words already exists in the frequency-of-occurrence storage 4 (S5). If it does not exist, the contents of words are memorized as a count 1 of an appearance to the frequency-of-occurrence storage 4 (S6), only 1 counts up Variable j (S7), and it returns to said step S2.

[0015] In said step S5, if the word train in words has already existed in the frequency-of-occurrence storage 4, only 1 will count up the count of an appearance of the contents of words memorized by the frequency-of-occurrence storage 4 (S8), and it will go to said step S7. In said step S2, if the value whose value of Variable j is the number N of the maximum copula connection is exceeded, 1 will be set to j, i will be counted up one time, and it will go to (S9) and said step S3. In said step S4, it judges whether if the word of eye watch (i+j -1) does not exist next, j is equal to i (S10), and if not equal, if equal, i-1 will be set to going and the variable A which memorizes the total number of words of an object document to said step S9 (S11).

[0016] Next, whenever [coincidence / of the copula of the amount size N individual memorized by the frequency-of-occurrence storage 4] is calculated. The information which memorized the result to the frequency-of-occurrence storage 4 (S12), and was memorized by the frequency-of-occurrence storage 4 with classification equipment 6 is changed together with the high order of whenever [coincidence] (S13).

[0017] Hereafter, the example 1 of this invention is concretely explained based on an example. The text is inputted from the object document input section.

[0018]

[Table 1]

The orchestra gave him superb support. ←—— 入力文

(語番号1 2 3 4 5 6)

[0019] The example which carries out automatic extracting of the 3 grams [a maximum of] copula is explained. First, initial value 1 is set to one variables i and j (S1). Since an object document is read from the object input section 1 and Variable j is not over the three maximum copulas (S2), one word (j= 1) is obtained from a head (i= 1) (S3). That is, "The" is obtained. Since an i+j-1=1 position word exists, (S4) and this word investigate whether the N-gram word storage 4 memorizes (S5). Since it does not memorize yet, a word "The" is newly memorized as the count 1 of an appearance to the N-gram word storage 4 (S6). When having already memorized, only 1 counts up the count of an appearance (S8). Only 1 counts up j (S7) and then two words (j+1) are obtained from a head (i= 1).

[0020] That is, "The orchestra" is obtained. Since an i+j-1=2 position word exists, (S4) and this word investigate whether the N-gram word storage 4 memorizes. Since it does not memorize yet, it memorizes newly and the count of an appearance is set to 1. At the time of j= 3, "The orchestra gave" is obtained similarly. It will be set to j= 4 if one j is counted up (S7). Since j exceeds three maximum copula connection, 1 is set to j, i is counted up one time, and it is referred to as 2 (S9).

[0021] Next, one word (j= 1) is obtained from the 2nd word (i= 2). "orchestra" is obtained. If the connection word is started like the case of i= 1, counting up j to a maximum of 3, "orchestra gave" and "orchestra gave him" will be started, and N-gram word storage will memorize with the count of an appearance. Said processing is repeated to i= 6. Finally the contents of drawing 4 are memorized by the N-gram word storage 4. Then, according to a formula (1), whenever [coincidence / of each copula] is calculated from the frequency of occurrence, and a result is memorized to the word storage 4 (S12). Furthermore, it does again a classification (sort) in the expensive order of whenever [coincidence].

[0022] For example, when preparing 1 million object input statements and extracting the connection word of a maximum of 3, the N-gram word storage 4 after count becomes like drawing 5 whenever [coincidence]. It is not asking for whenever [coincidence] about the case of N= 1. It sorts in the magnitude of the value of whenever [coincidence], for example, a threshold, such as making 50% of high orders into a candidate, is decided, or the number of copulas sorts with the value of whenever [coincidence] to things same (as the sort approach). There can be quick sort, a merge sort, a simple sorting method, etc., and can carry out, and the method of setting up the value of whenever [coincidence / of the copula extracted for every copula of each die length] more than with a certain threshold can extract the strong copula of whenever [coincidence] automatically.

[0023] The part which drawing 6 is a block diagram for explaining other examples (example 2) of the copula automatic extracting equipment by this invention, and seven are conditioning storage pattern match equipment and 8 among drawing, in addition carries out the same operation as drawing 1 has attached the same sign.

[0024] N-gram (N= 1, 2 and 3, --, Nmax) of a word is started with N-gram logging equipment 2 from the object document which read the object document from the object document input section 1, and was read from this object document input section 1. The frequency of occurrence of the connection word of N-gram started by this N-gram logging equipment 2 is totaled with frequency total equipment 3, and the frequency of occurrence totaled by said frequency total equipment 3 is remembered to be a N-gram connection word with the word storage 4.

[0025] Whenever [coincidence / of N-gram] is calculated with count equipment 5 whenever [coincidence] using the frequency of occurrence in the object document of each word which constitutes said N-gram (in the case of N= 1), and

the frequency of occurrence of N-gram itself. Classification equipment 6 puts in order and changes the information in the word storage 4 with the value of whenever [coincidence / which was calculated with count equipment 5 whenever / said coincidence]. Thus, it can discern whether they are whether the entry which consists of two or more words is coincided also to the idiom and copula by combination of the word which is not a high word of an expert, and the connection depended by chance, and the efficient degree of coincidence can collect strong copulas automatically.

[0026] The configuration which does not suit the conditions which should be extracted as vocabulary is memorized, pattern match equipment 7 is for eliminating N-gram suitable for the configuration memorized in said conditioning store 8, and the conditioning store 8 can extract a copula with still more sufficient extent by using these. That is, in an example 2, the configuration which does not suit an extraction condition as vocabulary requires very the pattern which cannot serve as a copula easily. For example, there is two connection of a word different from the case of 2 connection words of an article and one certain word or pronouns (English his, my, your, their, them, him, etc.) etc.

[0027] The example of an example 2 is shown below. The example of Table 2 is the case where it is eliminated from the object of the extract vocabulary, when it is "the", "a", "an", "his" or, and -- limited the number of connection with 2, and the first word of 2 connection words in that case was remembered to be in equipment. A pattern match is performed by the string comparison with pattern match equipment 7.

[0028]

[Table 2]

連接数	先頭語
2	the
2	a
2	an
2	his
2	her
2	their
2	my
2	your

[0029] Moreover, it is the case where the copula whose 2nd word it is "the", "a" or, and "an" limited the number of connection with 2, and the first word of 2 connection words in that case was remembered to be in equipment in the example of Table 3, and is "in", "of", "with", --"from", or "to" is eliminated from the object of the extract vocabulary. A pattern match is performed by the string comparison with pattern match equipment 7.

[0030]

[Table 3]

連接数	先頭語	第2単語
2	the	in
	a	of
	an	with
		for
		on
		from
		to

[0031]

[Effect of the Invention] According to this invention, there is the following effectiveness so that clearly from the above explanation.

(1) Effectiveness corresponding to claim 1 : the word which appears as a copula with strong coincidence in an input statement with easy equipment can be quickly extracted efficiently automatically without using a dictionary etc., since it can ask for the strength [the configuration word of a copula] of coincidence from the frequency of occurrence of the configuration word, and the frequency of occurrence of a copula.

(2) Effectiveness corresponding to claim 2 : since the N-gram pattern considered to be unsuitable for extracting as vocabulary in the word extracted above (1) is beforehand memorized to the conditioning store, what matches this pattern from a vocabulary candidate can be eliminated, and the vocabulary can be extracted with a sufficient precision.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL FIELD

[Industrial Application] This invention relates to the copula automatic extracting equipment for carrying out automatic collection of the copula efficiently from an object document at a detail more about the copula automatic extracting equipment in language-processing equipment. For example, it is applied to vocabulary dictionary listing devices, such as machine translation and a word processor.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art] As well-known reference which indicated conventional language-processing equipment, there is JP,6-19968,A, for example. In order for the thing of this official report to enable it to extract a technical term easily out of a huge word and to enable it to build a technical-term dictionary easily for a short time An input statement is divided into a word with word division equipment, and normalization of part-of-speech information being given is performed. The input data which it normalized with word division equipment is outputted to technical-term judging equipment, while this technical-term judging equipment refers said each dictionary, evaluation of each word is performed, and the candidate of a technical term is extracted according to this evaluation. However, a technical-term judging is performed in consideration of the number of configuration words, the operating frequency of a configuration word, the vocabulary dictionary classified by field, and a type of letters (katakana word), and the vocabulary dictionary classified by field is needed. Moreover, the thing of said official report does not have the description about the technical-term candidate selection for a judgment.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

EXAMPLE

[Example] An example is explained below with reference to a drawing. First, the connection word of N-gram is extracted from an object document (N= 1, 2 and 3, --, Nmax). If it is the case where an object document is English, it will refer to a space character etc. from the gestalt-description of language, and every word will be divided. If it is 3-gram of N= 3, the vocabulary which carries out a maximum of 3 word connection will be started. Extracted N connection word carries out the count total of the frequency of occurrence with the equipment which totals the frequency of occurrence. Moreover, the frequency of occurrence for every word is counted, and it totals. This result is memorized by N-gram word storage.
[0008] After the total of the frequency of occurrence to an input-statement document finishes, count of whenever [coincidence] is calculated according to the following formulas to the connection word of N-gram in N-gram word storage. When the words of N connection, i.e., the configuration word of a copula, are w1, w2, w3, --, wN, respectively, the frequency of occurrence of the N connection word itself expresses [each frequency of occurrence] with H (w1), H (w2), --, H (w3) H (w1, w2, w3, --, wN). Moreover, the total number of words of an object input-statement document is set to A.
[0009]
[Equation 1]

$$\frac{P(w1, w2, w3, ..., wN)}{P(w1)\times P(w2)\times P(w3)\times ... \times P(wN)} \quad \cdots (1)$$

ここで、

$$P(w1)= \frac{H(w1)}{A}$$

$$P(w2)= \frac{H(w2)}{A}$$

$$P(w3)= \frac{H(w3)}{A}$$

. . . . . .

$$P(wN)= \frac{H(wN)}{A}$$

$$P(w1, w2, w3, ..., wN) = \frac{H(w1,w2,w3,...wN)}{A-(N-1)} \quad \cdots (2)$$

である。

Aの値がNに比べて充分大きいとき、式（2）は $\frac{H(w1,w2,w3,...wN)}{A}$

と近似できる。

[0010] (1) The denominator of a formula expresses the probability which each word connects by chance from the appearance probability of each word which constitutes a copula. (1) The molecule of a formula is the probability for each word to be connected and to actually appear. Therefore, (1) type serves as a ratio of the probability which a certain copula actually connects, and the probability connected [chance]. (1) It can be said that the connection word of the N-

gram has the high degree which coincides and appears, so that the value of a formula is high. conversely, when low, it coincides -- as -- ** -- possibility of connecting by chance is high.

[0011] Drawing 1 is a block diagram for explaining one example (example 1) of the copula automatic extracting equipment by this invention, and, for N-gram logging equipment and 3, as for N-gram word storage (frequency-of-occurrence storage) and 5, the frequency total equipment of N-gram and 4 are [one / the object document input section and 2 / count equipment and 6] classification equipment whenever [coincidence] among drawing.

[0012] N-gram (N= 1, 2 and 3, --, Nmax) of a word is started with N-gram logging equipment 2 from the object document which read the object document from the object document input section 1, and was read from this object document input section 1. The frequency of occurrence of the connection word of N-gram started by this N-gram logging equipment 2 is totaled with frequency total equipment 3, and the frequency of occurrence totaled by said frequency total equipment 3 is remembered to be a N-gram connection word with the word storage 4.

[0013] Whenever [coincidence / of N-gram] is calculated with count equipment 5 whenever [coincidence] using the frequency of occurrence in the object document of each word which constitutes said N-gram (in the case of N= 1), and the frequency of occurrence of N-gram itself. Classification equipment 6 puts in order and changes the information in the word storage 4 with the value of whenever [coincidence / which was calculated with count equipment 5 whenever / said coincidence]. Thus, it can discern whether they are whether the entry which consists of two or more words is coincided also to the idiom and copula by combination of the word which is not a high word of an expert, and the connection depended by chance, and the efficient degree of coincidence can collect strong copulas automatically.

[0014] Drawing 2 and drawing 3 are the flow charts for explaining actuation of the copula automatic extracting equipment by this invention. Hereafter, according to each step (S), it explains in order. First, Variable i and Variable j are set to 1 (S1), and it judges whether the value whose value of Variable j is the number N of the maximum copula connection was exceeded (S2). If it is not over the value of N next, from the object document input section 1, j word is inputted from i word eye from the head of the text, and it stores in Variable words (S3). Next, if it judges whether the word of eye watch (i+j -1) exists and (S4) and a word exist next, it will judge whether the word train in words already exists in the frequency-of-occurrence storage 4 (S5). If it does not exist, the contents of words are memorized as a count 1 of an appearance to the frequency-of-occurrence storage 4 (S6), only 1 counts up Variable j (S7), and it returns to said step S2.

[0015] In said step S5, if the word train in words has already existed in the frequency-of-occurrence storage 4, only 1 will count up the count of an appearance of the contents of words memorized by the frequency-of-occurrence storage 4 (S8), and it will go to said step S7. In said step S2, if the value whose value of Variable j is the number N of the maximum copula connection is exceeded, 1 will be set to j, i will be counted up one time, and it will go to (S9) and said step S3. In said step S4, it judges whether if the word of eye watch (i+j -1) does not exist next, j is equal to i (S10), and if not equal, if equal, i-1 will be set to going and the variable A which memorizes the total number of words of an object document to said step S9 (S11).

[0016] Next, whenever [coincidence / of the copula of the amount size N individual memorized by the frequency-of-occurrence storage 4] is calculated. The information which memorized the result to the frequency-of-occurrence storage 4 (S12), and was memorized by the frequency-of-occurrence storage 4 with classification equipment 6 is changed together with the high order of whenever [coincidence] (S13).

[0017] Hereafter, the example 1 of this invention is concretely explained based on an example. The text is inputted from the object document input section.

[0018]

[Table 1]

The orchestra gave him superb support. ←―― 入力文

(語番号1 2 3 4 5 6)

[0019] The example which carries out automatic extracting of the 3 grams [a maximum of] copula is explained. First, initial value 1 is set to one variables i and j (S1). Since an object document is read from the object input section 1 and Variable j is not over the three maximum copulas (S2), one word (j= 1) is obtained from a head (i= 1) (S3). That is, "The" is obtained. Since an i+j-1=1 position word exists, (S4) and this word investigate whether the N-gram word storage 4 memorizes (S5). Since it does not memorize yet, a word "The" is newly memorized as the count 1 of an appearance to the N-gram word storage 4 (S6). When having already memorized, only 1 counts up the count of an appearance (S8). Only 1 counts up j (S7) and then two words (j+1) are obtained from a head (i= 1).

[0020] That is, "The orchestra" is obtained. Since an i+j-1=2 position word exists, (S4) and this word investigate whether the N-gram word storage 4 memorizes. Since it does not memorize yet, it memorizes newly and the count of an appearance is set to 1. At the time of j= 3, "The orchestra gave" is obtained similarly. It will be set to j= 4 if one j is

counted up (S7). Since j exceeds three maximum copula connection, 1 is set to j, i is counted up one time, and it is referred to as 2 (S9).

[0021] Next, one word (j= 1) is obtained from the 2nd word (i= 2). "orchestra" is obtained. If the connection word is started like the case of i= 1, counting up j to a maximum of 3, "orchestra gave" and "orchestra gave him" will be started, and N-gram word storage will memorize with the count of an appearance. Said processing is repeated to i= 6. Finally the contents of drawing 4 are memorized by the N-gram word storage 4. Then, according to a formula (1), whenever [coincidence / of each copula] is calculated from the frequency of occurrence, and a result is memorized to the word storage 4 (S12). Furthermore, it does again a classification (sort) in the expensive order of whenever [coincidence].

[0022] For example, when preparing 1 million object input statements and extracting the connection word of a maximum of 3, the N-gram word storage 4 after count becomes like drawing 5 whenever [coincidence]. It is not asking for whenever [coincidence] about the case of N= 1. It sorts in the magnitude of the value of whenever [coincidence], for example, a threshold, such as making 50% of high orders into a candidate, is decided, or the number of copulas sorts with the value of whenever [coincidence] to things same (as the sort approach). There can be quick sort, a merge sort, a simple sorting method, etc., and can carry out, and the method of setting up the value of whenever [coincidence / of the copula extracted for every copula of each die length] more than with a certain threshold can extract the strong copula of whenever [coincidence] automatically.

[0023] The part which drawing 6 is a block diagram for explaining other examples (example 2) of the copula automatic extracting equipment by this invention, and seven are conditioning storage pattern match equipment and 8 among drawing, in addition carries out the same operation as drawing 1 has attached the same sign.

[0024] N-gram (N= 1, 2 and 3, --, Nmax) of a word is started with N-gram logging equipment 2 from the object document which read the object document from the object document input section 1, and was read from this object document input section 1. The frequency of occurrence of the connection word of N-gram started by this N-gram logging equipment 2 is totaled with frequency total equipment 3, and the frequency of occurrence totaled by said frequency total equipment 3 is remembered to be a N-gram connection word with the word storage 4.

[0025] Whenever [coincidence / of N-gram] is calculated with count equipment 5 whenever [coincidence] using the frequency of occurrence in the object document of each word which constitutes said N-gram (in the case of N= 1), and the frequency of occurrence of N-gram itself. Classification equipment 6 puts in order and changes the information in the word storage 4 with the value of whenever [coincidence / which was calculated with count equipment 5 whenever / said coincidence]. Thus, it can discern whether they are whether the entry which consists of two or more words is coincided also to the idiom and copula by combination of the word which is not a high word of an expert, and the connection depended by chance, and the efficient degree of coincidence can collect strong copulas automatically.

[0026] The configuration which does not suit the conditions which should be extracted as vocabulary is memorized, pattern match equipment 7 is for eliminating N-gram suitable for the configuration memorized in said conditioning store 8, and the conditioning store 8 can extract a copula with still more sufficient extent by using these. That is, in an example 2, the configuration which does not suit an extraction condition as vocabulary requires very the pattern which cannot serve as a copula easily. For example, there is two connection of a word different from the case of 2 connection words of an article and one certain word or pronouns (English his, my, your, their, them, him, etc.) etc.

[0027] The example of an example 2 is shown below. The example of Table 2 is the case where it is eliminated from the object of the extract vocabulary, when it is "the", "a", "an", "his" or, and -- limited the number of connection with 2, and the first word of 2 connection words in that case was remembered to be in equipment. A pattern match is performed by the string comparison with pattern match equipment 7.

[0028]

[Table 2]

連接数	先頭語
2	the
2	a
2	an
2	his
2	her
2	their
2	my
2	your

[0029] Moreover, it is the case where the copula whose 2nd word it is "the", "a" or, and "an" limited the number of connection with 2, and the first word of 2 connection words in that case was remembered to be in equipment in the

example of Table 3, and is "in", "of", "with", --"from", or "to" is eliminated from the object of the extract vocabulary. A pattern match is performed by the string comparison with pattern match equipment 7.

[0030]

[Table 3]

連接数	先頭語	第2単語
2	the	in
	a	of
	an	with
		for
		on
		from
		to

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is a block diagram for explaining one example of the copula automatic extracting equipment by this invention.
[Drawing 2] It is a flow chart (the 1) for explaining actuation of the copula automatic extracting equipment by this invention.
[Drawing 3] It is a flow chart (the 2) for explaining actuation of the copula automatic extracting equipment by this invention.
[Drawing 4] It is drawing showing the example of storage of the N-gram word storage in this invention.
[Drawing 5] It is drawing showing other examples of storage of the N-gram word storage in this invention.
[Drawing 6] It is a block diagram for explaining other examples of the copula automatic extracting equipment by this invention.
[Description of Notations]
1--1 is [-- It is count equipment and 6 whenever / coincidence /. / -- It is classification equipment and 7. / -- It is pattern match equipment and 8. / -- It is a conditioning store.] the object document input section and 2. -- It is N-gram logging equipment and 3. -- It is the frequency total equipment of N-gram, and 4. -- It is a N-gram word store and 5.

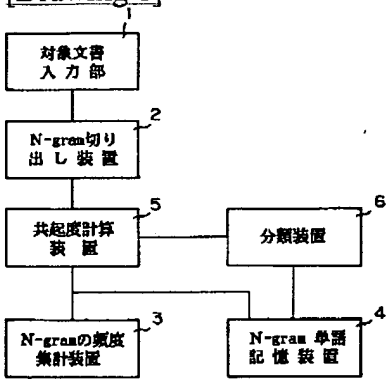
[Translation done.]

* NOTICES *

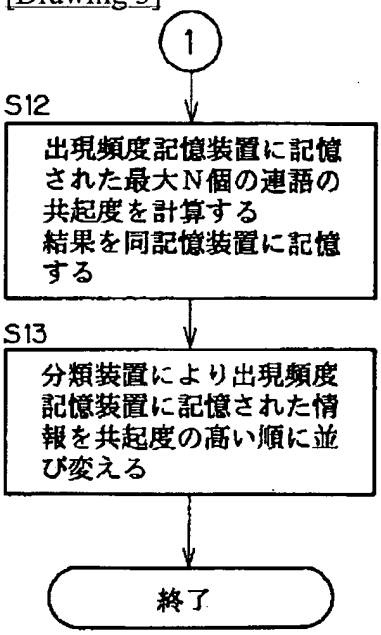
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

[Drawing 1]



[Drawing 3]



[Drawing 4]

連語	出現頻度	共起度
The	1	-
The orchestra	1	-
The orchestra gave	1	-
orchestra	1	-
orchestra gave	1	-
orchestra gave him	1	-
gave	1	-
gave him	1	-
gave him superb	1	-
him	1	-
him superb	1	-
him superb support	1	-
superb	1	-
superb support	1	-
support	1	-

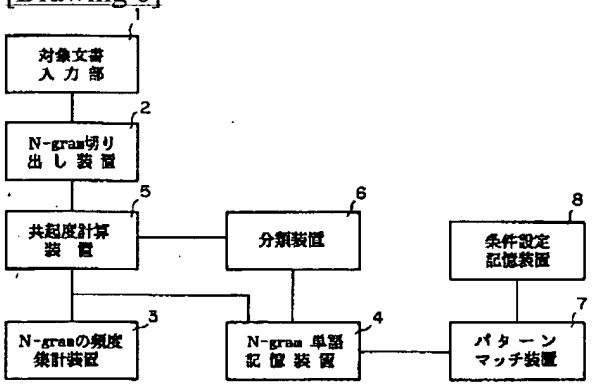
N-gram単語記憶装置の記憶例

[Drawing 5]

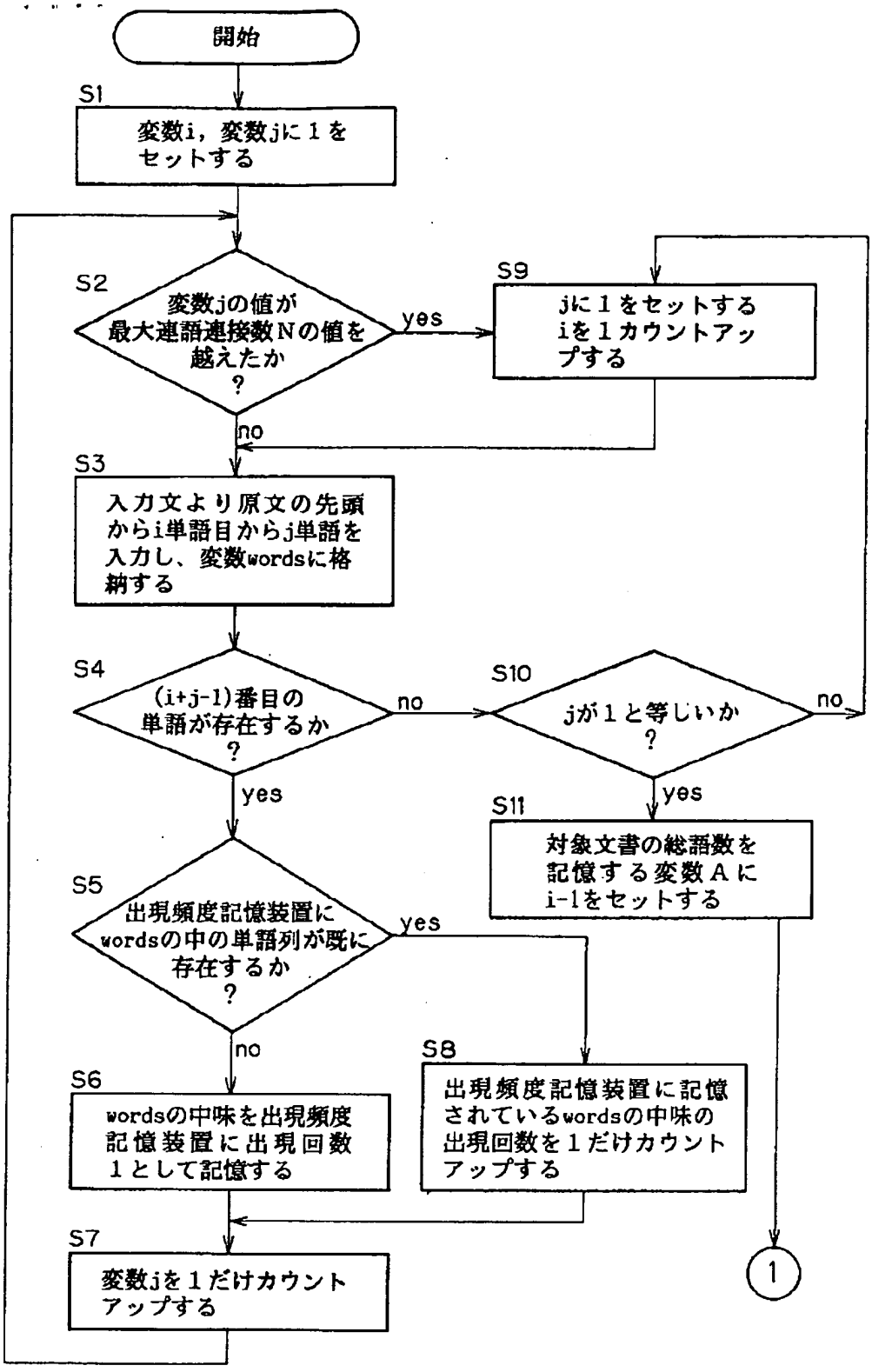
N連語	出現頻度	共起度
securities brings fannie	5	396920
securities	1654	
brings	64	
fannie	119	
the khmer rouge	12	306784
the	116415	
khmer	14	
rouge	24	
calls seeking comment	10	138985
calls	335	
seeking	342	
comment	628	
premium over yesterday's	6	67825
premium	165	
over	2437	
yesterday's	220	
public employees retirement	6	47793
public	1082	
employees	586	
retirement	198	
securities brings	8	758
brings fannie	8	1050
the khmer	12	7
khmer rouge	14	41667
calls seeking	11	96
seeking comment	11	51
premium over	28	70
over yesterday's	8	8
public employees	10	15
employees retirement	8	69

N-gram単語記憶装置の他の記憶例

[Drawing 6]



[Drawing 2]



[Translation done.]